Raymond Lin
304937942

CS M146 - Problem Set 1

1.a) For $n=4$, mistakes $=2$
For $n=5$, mistakes $=4$
For $n=6$, mistakes $=8$

For $n \geq 4$, mistakes $= \frac{2^n}{2^3} = 2^{n-3}$

b) No. For each $X \in \{X_4 \dots X_n\}$, the probability of X being a mistake is always $\frac{1}{8}$, which is the same as if we had a decision tree with 0 internal nodes. For each $X \in \{X_1, \dots X_3\}$, the probability of X being a mistake is $> \frac{1}{8}$. Thus, no split on any feature would reduce the number of mistakes.

c) 1-leaf decision tree $\Rightarrow$ always choose 1

$$H(Y) = \sum_k P(k) \cdot \log(P(k))$$

$$H(Y) = \sum_k P(k) \cdot \log(P(k))$$

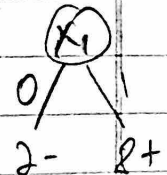
$$= P(\text{incorrect}) \cdot \log(P(\text{incorrect})) + P(\text{correct}) \cdot \log(P(\text{correct}))$$

$$= \frac{2^{n-3}}{2^n} \cdot \log\left(\frac{2^{n-3}}{2^n}\right) + \frac{2^n - 2^{n-3}}{2^n} \cdot \log\left(\frac{2^n - 2^{n-3}}{2^n}\right)$$

$$= \frac{2^n \cdot 2^{-3}}{2^n} \cdot \log\left(\frac{2^n \cdot 2^{-3}}{2^n}\right) + \frac{2^n\left(\frac{7}{8}\right)}{2^n} \cdot \log\left(\frac{2^n\left(\frac{7}{8}\right)}{2^n}\right)$$

$$= \frac{1}{8} \cdot \log\left(\frac{1}{8}\right) + \frac{7}{8} \cdot \log\left(\frac{7}{8}\right)$$

$H(Y) = 0.54$



d) Try split by X_1, X_2, or X_3

$$H(Y|X_1) = -\left(P(X_1=1) \cdot H(Y|X_1=1) + P(X_1=0) \cdot H(Y|X_1=0)\right)$$

$$H(Y|X_1=1) = -\left(P(Y=1|X_1=1) \cdot \log(P(Y=1|X_1=1)) + P(Y=0|X_1=1) \cdot \log(P(Y=0|X_1=1))\right)$$

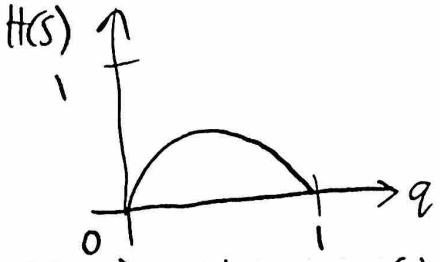
$$P(Y=1|X_1=1) = 1$$
$$P(Y=0|X_1=1) = 0$$
$$H(Y|X_1=1) = -(1 \cdot \log(1) + 0 \cdot \log(0)) = 0$$
$$H(Y|X_1=0) = -(P(Y=1|X_1=0) \cdot \log(P(Y=1|X_1=0)) + P(Y=0|X_1=0) \cdot \log(P(Y=0|X_1=0)))$$
$$P(Y=1|X_1=0) = \frac{3}{4}$$
$$P(Y=0|X_1=0) = \frac{1}{4}$$
$$H(Y|X_1=0) = -(\frac{3}{4} \cdot \log(\frac{3}{4}) + \frac{1}{4} \cdot \log(\frac{1}{4}))$$

$$P(X_1=1) = \frac{1}{2}$$
$$P(X_1=0) = \frac{1}{2}$$
$$H(Y|X_1) = (P(X_1=1) \cdot H(Y|X_1=1) + P(X_1=0) \cdot H(Y|X_1=0)$$
$$= (\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot (-(\frac{3}{4} \cdot \log(\frac{3}{4}) + \frac{1}{4} \cdot \log(\frac{1}{4})))$$
$$= \frac{1}{2}(-\frac{3}{4} \cdot \log(\frac{3}{4}) - \frac{1}{4} \cdot \log(\frac{1}{4}))$$

$$\approx 0.406$$

$$\boxed{H(Y|X_1) = 0.406}$$

Graph



2. a) $0 \leq H(S) \leq 1$, $H(S)=1$ when $p=n$

$$H(S) = B(q) = B\left(\frac{p}{p+n}\right)$$
$$= -\left(\frac{p}{p+n}\right)\log\left(\frac{p}{p+n}\right) - \left(1-\left(\frac{p}{p+n}\right)\right)\log\left(1-\left(\frac{p}{p+n}\right)\right)$$
$$= \left(-\frac{p}{p+n}\right)\log\left(\frac{p}{p+n}\right) - \left(\frac{p+n-p}{p+n}\right)\log\left(\frac{p+n-p}{p+n}\right)$$

- we know that p and n refer to a number of a sample, so p and $n \geq 0$
- let $p=0$,

$$H(S) = \left(-\frac{0}{n}\right)\log\left(\frac{0}{n}\right) - \left(\frac{n}{n}\right)\log\left(\frac{n}{n}\right)$$
$$= 0 \cdot \log(0) - 1 \cdot \log(1)$$
$$= 0 - 0 = 0$$

- likewise, when $n=0$, $H(S)=0$

- since p and $n \geq 0$, we know $0 \leq \frac{p}{p+n} \leq 1$ and $0 \leq \frac{n}{p+n} \leq 1$
- then $|\log(x)| < 1$ when $0 \leq x \leq 1$

$$\left|\log\left(\frac{p}{p+n}\right)\right| < 1 \quad \text{and} \quad \left|\log\left(\frac{n}{p+n}\right)\right| < 1$$

- we also know, $\frac{p}{p+n}$ is a compliment of $\frac{n}{p+n}$

$$1 - \frac{p}{p+n} = \frac{p+n-p}{p+n} = \frac{n}{p+n}$$

- Thus, $H(S) = \underbrace{\left(-\frac{p}{p+n}\right)}_{\text{compliment } \frac{n}{p+n}} \underbrace{\log\left(\frac{p}{p+n}\right)}_{-1 \leq x \leq 0} - \underbrace{\left(\frac{n}{p+n}\right)}_{\text{compliment } \frac{p}{p+n}} \underbrace{\log\left(\frac{n}{p+n}\right)}_{-1 \leq x \leq 0} \leq 1$

- $\therefore 0 \leq H(S) \leq 1$ □

- when $n=p$, we have

$$H(S) = \left(-\frac{p}{p+n}\right)\log\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right)\log\left(\frac{n}{p+n}\right)$$

$$= \left(-\frac{n}{2n}\right)\log\left(\frac{n}{2n}\right) - \left(\frac{n}{2n}\right)\log\left(\frac{n}{2n}\right)$$

$$= \left(-\frac{1}{2}\right)\log\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right)$$

$$= -\log\left(\frac{1}{2}\right)$$

$$= -\log(2^{-1})$$

$$= -(-1)\log(2)$$

$$= \log 2$$

- let base $= 2$, then

$$= \log_2 2$$

$$= 1$$

□

b) We have k disjoint subsets: $S \in \{S_1, S_2 \dots S_k\}$
Let the attribute we split on be $\underline{X}$.

$$GAIN(S, \underline{X}) = H(S) - H(S|\underline{X})$$

$$p = \sum_k p_k$$

$$n = \sum_k n_k$$

$$H(S) = B\left(\frac{p}{p+n}\right)$$

$$H(S|\underline{X}) = \sum_k \frac{p_k+n_k}{p+n} \cdot B\left(\frac{p_k}{p_k+n_k}\right)$$

$$= \frac{\sum p_k+n_k}{p+n} \cdot B\left(\frac{p}{p+n}\right)$$

$$= \frac{p_k+p_k \dots n_k+n_k}{p+n} \cdot B\left(\frac{p}{p+n}\right) = \frac{p+n}{p+n} \cdot B\left(\frac{p}{p+n}\right)$$

$$GAIN(S, \underline{X}) = H(S) - H(S|\underline{X}) = B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right) = 0$$ □

④

3.

a) when $k=1$, training set error is minimized. We can count the point itself as its own neighbor. Thus, training set error $=0$ when $k=1$.

This is not a reasonable estimate of test set error because we never get any error.

b)

K	O errors	* errors	total errors	error rate
1	5	5	10	71%
3	3	3	6	43%
5	2	2	4	29%
7	2	2	4	29%
9	7	7	14	100%
11	7	7	14	100%
13	7	7	14	100%

$k=5, 7$ minimize the LOOCV error for this data set. Error $\approx 29\%$. Cross validation is a good measure of test set performance because we we all parts of the dataset to train the model

c) $k=1$, error $= 71\%$
$k=13$, error $= 100\%$

Using too small a value, we overfit the data, because even outlier data points are considered.

Using too large a value, we underfit the data, because it will bias the more dominant label.

4.1

a)

The created histograms show the relationship between particular features and the number of people who survived the Titanic.

1. The first feature that is being shown is the class of the ticket, named “Pclass”. We can see that the higher class tickets had a higher survival rate, while the lower class tickets had a lower survival rate. For example, first class ticket holders had a higher survival rate than third class ticket holders.
2. The second features that is shown is the gender. While males and females are not labeled, we can assume from logic that more females survive than males.
3. The third feature that is shown is the age. We can see that very young children had a high survival rate, while middle aged people had low survival rates. Additionally, we can see the distribution of ages of passengers on the Titanic. The majority of passengers are middle aged, and there were fewer children and elderly.
4. The fourth feature is a particular individual’s number of siblings or spouses who are also aboard the Titanic. We can see that most people had no siblings aboard, and as the number of siblings decreased, the frequency of people decreased.
5. The fifth feature is a particular individual’s number of parents and children who are also aboard the Titanic. This feature follows the same trend as the siblings/spouses feature.
6. The sixth feature shows the fare price and the survival rates. The data shows that the cheaper the ticket, the lower the survival rate.
7. The seventh feature shows the port of embarkation or origin. It appears that most of the passengers embarked at Southampton, while fewer passengers embarked on Cherbourg and Queenstown. It is also worth noting that the survival rate for Cherbourg passengers was higher than that of the other two origins.

4.2

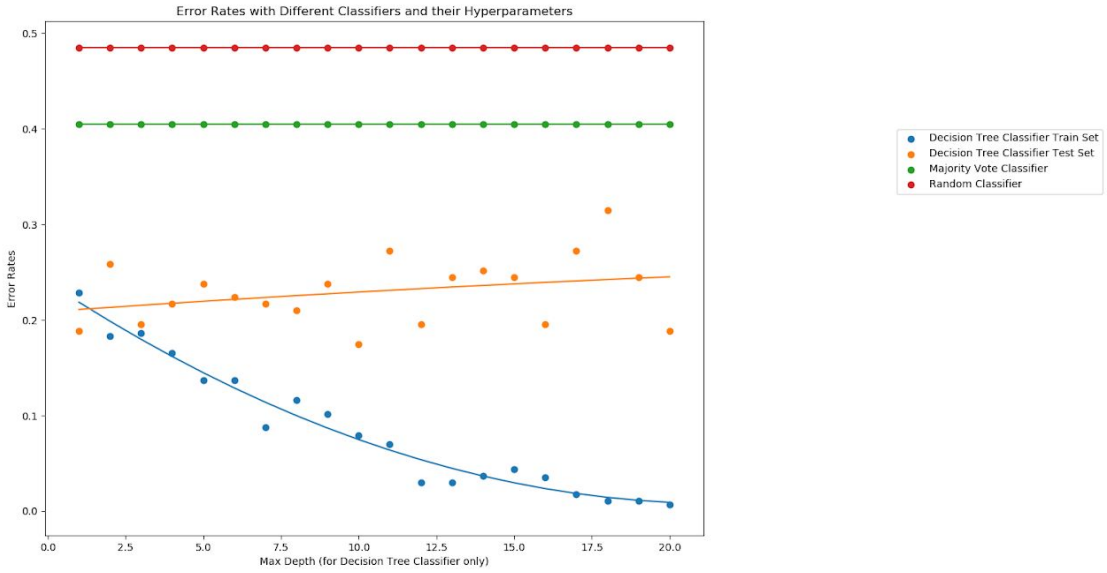
b) error = 0.485

c) decision_tree_error = 0.014

d)

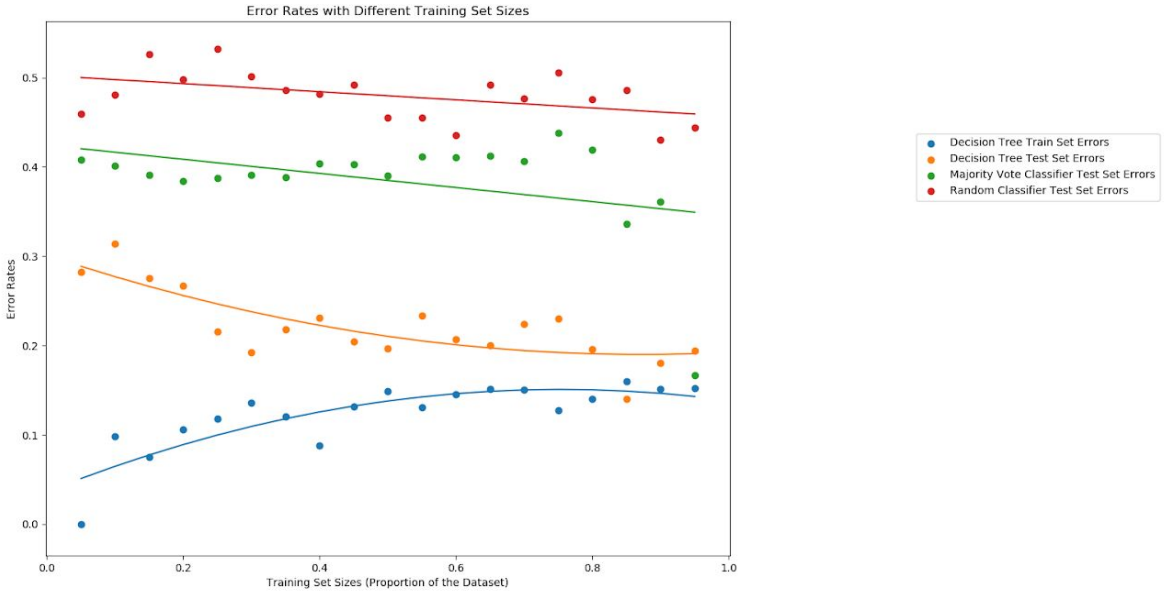
	Train Error	**Test Error**
Majority Vote Classifier	0.404	0.406
Random Classifier	0.527	0.498
Decision Tree Classifier	0.012	0.238

e)



The best depth limit for the decision tree classifier is approximately 5 levels. For the decision tree classifier, we want to achieve the best results without having a tree that is too deep. If the tree is allowed to be too deep, then overfitting may occur, which means that the data was memorized instead of learned. We also don't want the tree to be too shallow, because then there would not be enough learning done and error rates would remain high. In this case, overfitting does occur, because the test set error rate increases as the maximum depth hyperparameter increases.

f)



From the plot, we can see that the error rates for the decision tree classifier test set decrease as the training size increases. This is because as we increase the training set size, we learn more about the data, so our predictions become more accurate. In contrast, we can see that the error rates for the decision tree classifier for training set increases. This is because there is more data for the model to learn and thus it becomes more difficult for the model to accurately predict such a large dataset with high variance, especially since we have set a limit to its max depth parameter. For the random and majority vote classifiers, the error rates remain relatively stable, perhaps decreasing a little bit. This is consistent with our hypotheses, because we never change the proportion of survivors (in the majority vote classifier case) or the probability of survival (in the random classifier case), and thus the error rates remain roughly the same.